Peekaboo

- □□□□□PEEKABOO: Interactive Video Generation via Masked-Diffusion
- □□□□□https://arxiv.org/abs/2312.07509
- CVPR 2023

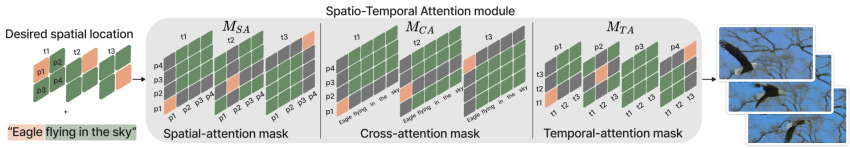

Figure 2. **PEEKABOO Module**: Our method proposes converting attention modules of an off-the-shelf 3D UNet into masked spatio-temporal mixed attention modules. We propose to use local context for generating individual objects and hence, guide the generation process using attention masks. For each of spatial-, cross-, and temporal-attentions, we compute attention masks such that foreground pixels and background pixels attend only within their own region. We illustrate these mask computations for an input mask (size 2×2 and 3 frames) which changes temporally as shown on the left. Green pixels are background pixels and orange are foreground. In the attention masks, both green and orange pixels have a value of 1, and gray pixels have a value of 0. We add the colors for ease of exposition. This masking is applied for a fixed number of steps, after which free generation is allowed. Hence, foreground and background pixels are hidden from each other before being visible, akin to a game of PEEKABOO. Best viewed in color.

□□□□□□training-free□□□bbox□□□□□□□□□□□□□□□□□□□□□□□□□□□bbox□□□□□□ attention□mask□□□□□U-Net□□□□□□□□□□self-attention mask□□□□cross-attention mask□□□box□□□□□□□□□□□□□□token□□box□□□

Figure 1. **Zero-training No-latency interactive video generation.** PEEKABOO allows users to control the output (object size, location and trajectory) for any off-the-shelf video diffusion models, through specially designed masking modules. First row shows a panda playing PEEKABOO by following an expanding mask in left direction.

- □□□□□□□□□□benchmark
- □□□mIoU; CD(□□□□□)□AP50□Cov□FVD
- □□□□□□
- □□□https://yash-jain.com/projects/Peekaboo/

Newer Older

2025-07-14
TV-LiVE

2025-07-14
Enhance-A-Video